图书情报工作
第 62 卷 第 21 期 2018 年 11 月

基于改进内容过滤算法的高校图书馆文献资源个性化推荐研究*

耿立校[1] 晋高杰[1] 李亚函[1] 孙卫忠[1,2] 马士豪[1]
[1] 河北工业大学经济管理学院 天津 300401 [2] 河北工业大学图书馆 天津 300401

摘要：[目的/意义]基于内容的过滤推荐中，针对向量空间模型表示文本时容易造成维度灾难的问题，提出利用余弦值 r 与匹配度值 Sim 相结合的方法对原有模型进行改进。[方法/过程]由文献资源和用户兴趣分别筛选出权重较大特征词的词向量，进而由公式计算余弦值 r，结合对应的特征词权重进一步计算出匹配度值 Sim，将其作为向目标用户推荐文献的依据，并利用河北工业大学图书馆的相关数据对改进模型、向量空间模型及 LDA 主题模型进行实验，最后利用查准率、召回率、F1 值及运行时间等评价指标对 3 种模型的实验结果进行分析。[结果/结论]实验结果表明所提出的改进模型相比较于实验中的向量空间模型与 LDA 主题模型具有更高的应用价值与运行效率。

关键词：基于内容推荐 匹配度值 Sim 推荐模型 实证分析
分类号：G252
DOI:10.13266/j.issn.0252-3116.2018.21.014

1 引言

电子文献已成为高校图书馆馆藏的重要组成部分，具有种类多且数据量不断增大的特点，用户在利用信息的过程中容易出现信息过载与信息迷航的现象，浪费很多时间和精力，却很难获得自己想要的文献[1-3]。随着高校科研用户研究方向的多元化，逐渐产生了个性化的信息需求，期望图书馆能够依据自己的兴趣爱好，得到经过筛选的电子文献。个性化推荐技术以用户浏览、收藏、下载等记录为基础，利用数据挖掘技术提取出用户的兴趣特征，基于用户兴趣特征在资源库中寻找目标用户感兴趣的文献，来完成个性化的信息推荐[4]。因此，针对不同的用户需求，提供个性化的信息服务成为当下高校图书馆信息服务研究的重要课题。

2 研究现状

个性化推荐服务由推荐系统来实现，推荐系统一般由资源、用户、推荐算法 3 个要素组成，推荐算法是推荐系统的核心[5]。推荐系统的核心推荐算法可分为协同过滤推荐、基于内容的过滤推荐以及混合推荐。协同过滤推荐需要分析用户历史评分情况，为目标用户寻找兴趣相似的邻居或者相似的资源进行推荐[1]；基于内容的过滤推荐通过分析用户感兴趣的资源和资源库中其他资源的相似度，选择相似度较高的资源为目标用户进行推荐[6-7]；混合推荐主要是融合以上两种推荐方法为目标用户进行推荐[8]。协同过滤推荐算法在现实生活中有广泛的应用，如电子商务、电影推荐等，但它最大的缺点是需要大量的用户历史评分，由于高校图书馆电子文献没有用户的评分以及评价等反馈信息，所以该算法不适于本文的推荐。基于内容的过滤推荐算法在文本领域的应用非常广泛，考虑到电子文献的特点，本文选择基于内容的过滤推荐算法来实现高校图书馆电子文献的推荐。

在国内的研究中，基于内容的过滤推荐算法，将文本资源作为主要对象。如徐勇等[9]将基于内容的过滤推荐算法应用到科技文献的推荐上，采用向量空间模型来描述用户兴趣特征和科技文献特征，比较用户兴趣特征与科技文献特征的相似度，将相似度较高的科技文献推荐给用户。吕学强等[10]在基于内容的过滤推荐基础上，提出了一种结合影评内容相似度和长短

* 本文系河北省社会科学基金项目“面向用户科研需求的高校图书馆信息服务体系研究”(项目编号:HB17TQ009)研究成果之一。
作者简介：耿立校(ORCID:0000-0002-1041-5061)，副教授，博士，硕士生导师，E-mail:lixgeng@qq.com；晋高杰(ORCID:0000-0002-0630-5986)，硕士研究生；李亚函(ORCID:0000-0003-2816-8596)，助理研究员，硕士；孙卫忠(ORCID:0000-0002-6073-7114)，馆长，教授，博士，硕士生导师；马士豪(ORCID:0000-0003-3250-795X)，硕士研究生。
收稿日期:2018-04-25 修回日期:2018-06-24 本文起止页码:112-117 本文责任编辑:王传清

期兴趣模型的方法来计算电影相似度，以此来对用户进行推荐。安悦等[11]针对微博信息过载的问题，将基于内容的过滤推荐算法应用到微话题的推荐上，通过计算微话题与用户兴趣的相似度，为微博用户推荐感兴趣的微话题。丁德红等[12]使用基于内容的过滤推荐，构建用户兴趣模型和文档特征模型，通过计算模型的相似度，将相似度较高的文档推荐给用户。雷凯等[13]利用基于内容的过滤推荐并融入实时交通路况，为用户推荐路线信息。

区别于国内学者以文本资源作为推荐对象的情况，国外一些学者将基于内容过滤的推荐算法应用在其它领域上。如 L. Liu 等[14]提出一种基于语义内容的推荐方法，将基于内容的推荐与上下文分析结合起来，在缺少用户反馈的条件下为用户提供软件推荐服务。Y. Deldjoo 等[15]利用基于内容的过滤推荐算法，结合视频视觉特征自动提取技术，为用户进行视频推荐，并与利用诸如电影流派等显性特征进行推荐的现有基于内容的推荐系统进行比较，证明前者有更高的推荐准确率。S. Lee 等[16]利用基于内容的过滤推荐算法，为韩国公开市场网购的用户提供可靠的卖家，首先将卖家划分为可信卖家与不可信卖家，然后使用基于内容的过滤推荐在选定的可信卖家中找到匹配度较高的前 K 个卖家。N. H. Liu [17]在基于内容的过滤推荐算法基础上，提出了一种基于用户偏好来计算不同乐曲之间个性化距离测量的方法，为用户进行个性化的音乐推荐服务。

综上所述，在基于内容的过滤推荐中，广泛采用基于向量空间模型(Vector space model, VSM)的推荐，计算特征词权重常用方法是 TF-IDF。但对于高校图书馆这样庞大的资源库来说，如果将电子文献表示成 VSM 的形式其维度将会非常大，易造成维度灾难，使得推荐系统效率低下。为避免维度灾难，区别于以往将文本资源和用户兴趣分别表示成空间向量模型的方法，进而计算出两个向量的相似度为用户进行资源推荐，本文提出余弦值 r 与匹配度值 Sim 相结合的方法对基于向量空间模型进行推荐的方法进行改进。其中 r 是指从文献资源与用户兴趣中筛选出的权重较大的特征词向量的余弦值，结合对应的特征词权重进一步计算出用户兴趣与文献资源的匹配度值 Sim，将 Sim 值较高的文献推荐给目标用户。相比较于原有基于向量空间模型的推荐：一方面，筛选出权重较大者作为最终的特征词，既能很好的代表文献资源特征，避免文本向量化时带来维度灾难的问题，又降低了计算的复杂度，提高推荐模型的计算效率；另一方面，从用户角度看，筛选出权重较大的特征词可以更准确的表达用户的兴趣，避免由于用户兴趣的泛化导致推荐范围过于宽泛，而失去个性化推荐的意义。最后通过实证分析来验证改进模型的有效性，以期能够为高校图书馆个性化推荐服务提供新的思路。

3 个性化推荐模型的构建

在以往基于内容的过滤推荐中，使用向量空间模型来表示文献资源与用户兴趣，即利用提取出的文献特征词及其权重来表示文献资源特征，用户兴趣特征主要是以兴趣信息提取出的特征词及其权重来表示，最后计算文献资源特征与用户兴趣特征的相似度，将相似度较高的文献推荐给目标用户。本研究在构建个性化推荐模型时将直接利用提取出的特征词及其权重、训练出的特征词向量来实现用户兴趣与文献资源的匹配度计算，首先利用公式 1 计算余弦值 r。

$$r = \frac{|u_i \cdot v_j|}{|u_i| \cdot |v_j|} \quad \text{（公式 1）}$$

其中 u_i 表示文献筛选出的第 i 个权重较大的特征词向量，v_j 表示用户兴趣筛选出的第 j 个权重较大的特征词向量，其中 $1 \leq i \leq P, 1 \leq j \leq Q$，P、Q 分别为文献与用户兴趣筛选出的权重较大的特征词数量。

然后利用公式 2 计算匹配度值 Sim。

$$Sim = \sum_{i=1}^{P} \sum_{j=1}^{Q} p_i \cdot q_j \cdot r = \sum_{i=1}^{P} \sum_{j=1}^{Q} \frac{p_i\, q_j |u_i \cdot v_j|}{|u_i| \cdot |v_j|} \quad \text{（公式 2）}$$

其中，p_i、q_j 分别表示从文献与用户兴趣中筛选出的权重较大的特征词权重，$1 \leq i \leq P, 1 \leq j \leq Q$。

个性化推荐模型主要有 3 个模块：文献处理模块、用户兴趣处理模块、文献推荐依据计算模块。其中文献处理模块是用于文献特征提取；用户兴趣处理模块是用于用户兴趣特征提取；文献推荐依据计算模块用来计算文献资源与用户兴趣的匹配度值 Sim，该模块是本文模型改进的核心部分，见图 1。

由图 1 可见，个性化推荐模型主要内容包括：

(1) 数据预处理。对搜集的用户阅读记录信息和文献进行数据清洗、文本类型转换等，使其符合模型对数据的要求，以便于分词、特征提取及词向量训练。

(2) 分词。分词通过分词器来实现，分词器提取关键词的准确性对提高推荐准确率有很大的影响。本研究选择目前最好的 Python 类中文分词器 jieba 分词，它采用目前流行的 Python 语言进行编码，支持繁体中文

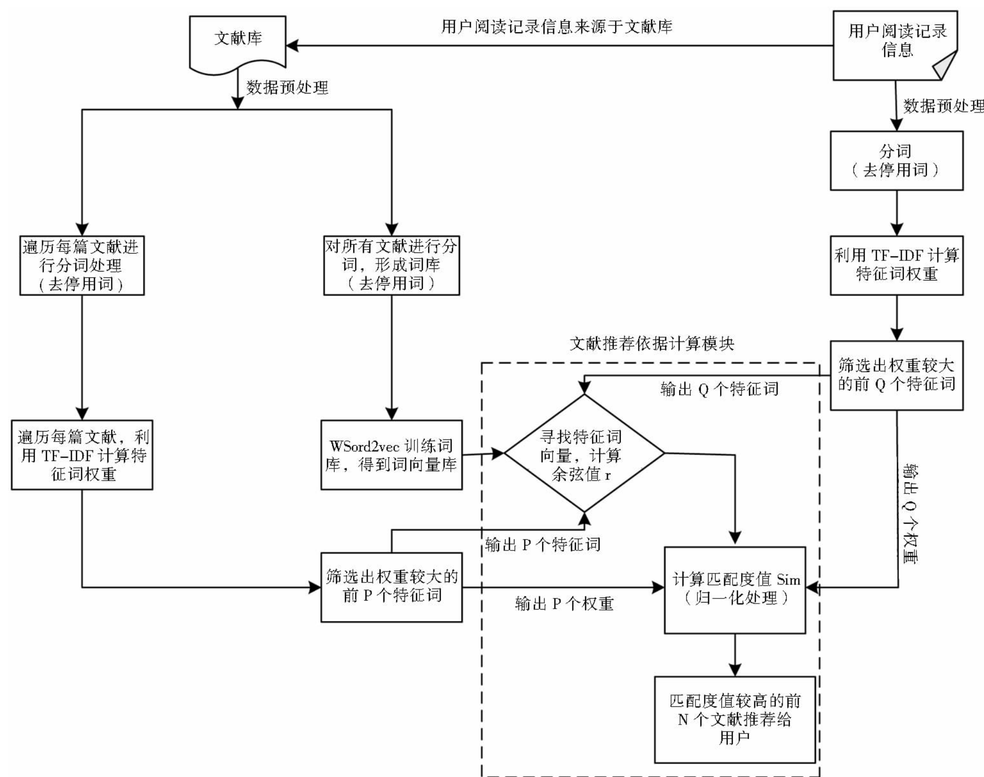


图 1　个性化推荐模型

词及自定义词典，还提供了精简模式、全模式、搜索模式等分词模式满足不同用户需求。

（3）TF-IDF。利用 TF-IDF 算法对用户兴趣信息与文献进行特征提取，该算法主要思想是：一个词在特定的文档中出现的频率越高，在所有文档中出现的范围越小，说明该词在区分文档内容属性方面的能力越强，特征词的权重计算公式为 $W = TF * IDF$[18]。

（4）Word2vec。利用 Word2vec 对分词后的文献库进行词向量训练，为计算特征词向量的余弦值做准备。该工具提供了 CBOW 和 Skip_gram 两种训练模型[19]，结合 hierarchy softmax 与 negative sampling 的优化技术，word2vec 可以高效的将词语表达成向量[20]。

（5）文献推荐依据计算模块。该模块是改进模型的核心内容，余弦值 r 与匹配度值 Sim 依据公式 1 与公式 2 得出，对 Sim 值进行归一化处理后，将 Sim 值较大的前 N 个文献推荐给目标用户。

个性化推荐模型的流程顺序为：①依次遍历文献库中的每篇文献，对其进行预处理，然后分词、去停用词，利用 TF-IDF 技术计算每篇文献的特征词权重，最后筛选出权重较大的前 P 个特征词。②对文献库中的所有文献进行预处理，然后分词、去停用词形成词库，利用 Word2vec 技术对词库进行训练，得到特征词向量库。③对用户阅读记录信息进行预处理，然后分词，去停用词，利用 TF-IDF 技术计算用户的特征词权重，并筛选权重较大的前 Q 个特征词。④在①、②、③都完成的情况下，利用筛选出的 P 与 Q 在②训练出的特征词向量库中寻找对应的特征词向量，利用公式 1 计算出余弦值 r。⑤利用公式 2 计算出文献与用户兴趣的匹配度值 Sim（归一化处理），把 Sim 值较大的前 N 个文献推荐给目标用户。

4　实证分析

为了验证改进模型的有效性，本研究使用河北工业大学图书馆的相关数据，计算改进模型下用户平均文献推荐准确率，并探索所提出的算法中 P 与 Q 的最优取值。最后在相同的实验环境及数据条件下，对改进模型、向量空间模型及 LDA 主题模型进行实验，分析 3 种推荐模型的查准率 P、召回率 R、F1 值及运行时间等评价指标。

4.1　数据来源

实验选取 8 名不同研究主题用户阅读过的文献作为实验数据，选取研究主题时遵循两个原则：①研究主题的完全无关性；②研究主题的相似性，研究主题涉及环境保护、健康医疗、建筑设计、数学、化学、物理、机械工程以及电气工程。依据人工分类将文献分为用户阅读记录文献与相关文献，其中用来提取用户兴趣特征

的阅读记录文献173篇，相关文献240篇。为了使实验更加符合实际情况，添加其他研究主题的干扰文献82篇，文献数量共计495篇，如表1所示：

表1　实验数据一览

用户	研究主题	阅读记录文献(篇)	相关文献(篇)
User1	环境保护	19	30
User2	健康医疗	20	30
User3	建筑设计	20	30
User4	数学	20	30
User5	化学	20	30
User6	物理	19	30
User7	机械工程	15	30
User8	电气工程	20	30
干扰文献	其他	0	82
合计	–	173	322

4.2　实验过程

(1)实验环境。在Windows系统环境下使用Python(3.5)对文本数据进行分析。分别使用jieba(0.39)、gensim(1.2.1)对文本进行分词及词向量训练。

(2)参数设置。Word2vec训练参数参考相关文献进行设置[21－22]。其中size＝100,window＝5,min_count＝5,sample＝1e－3,sg＝0,hs＝0,negative＝5,cbow_mean＝1。此外，由于用户相关文献数量的最大值为30，则令文献推荐数量top_n＝30。

(3)实验步骤。包括:①将所有TXT格式的文献(包括阅读记录文献、相关文献、干扰文献)放入文件夹A。②将User1的TXT格式的阅读记录文献放入另一个文件夹B。③对文件夹A内的每篇文献遍历并用jieba分词器单独分词，得到每篇文献语料库a_t(t＝1、2、3、…、495)。④合并语料库a_1、a_2、a_3、…、a_{495}形成总语料库D。⑤遍历语料库a_1、a_2、a_3、…、a_{495}，使用TF-IDF计算语料a_t的特征词权重，筛选权重较大的前P个特征词，形成文献t的特征词库$F_t=(F_{t1}、F_{t2}、\cdots、F_{tp})$，相应的权重$w_t=(W_{t1}、W_{t2}、\cdots、W_{tp})$。⑥使用gensim中的word2vec对总语料库D训练词向量，形成词向量库V。⑦提取User1阅读记录文献中的特征词，计算每个特征词权重的平均值，并筛选平均权重较大的前Q个特征词作为User1兴趣特征词，记为FI，相应的权重集合记为WI。⑧依次遍历所有文档，计算所有F_t与FI组合所对应V中特征词向量的余弦值，即依据公式1计算特征词向量的余弦值。⑨依据公式2计算User1与每篇文献的匹配度值Sim。⑩将阅读记录文献从Sim值列表中剔除，并将剩余文献按Sim值大小排序，按照公式计算User1的推荐准确率。⑪更新文件夹B中的阅读记录文献，分别导入User2、…、User8阅读记录文献，再次运行上述过程。

上述实验步骤中P与Q的值通过下面的实验来确定。首先令P＝5,Q依次取5、10、15、20、25、30、35、40、45、50，得到如图2所示的8名用户的平均推荐准确率。由2图可见，随着Q值的不断增大，平均推荐准确率逐渐提高，当Q为40、50时平均推荐准确率不再发生明显变化，因此取Q＝40。

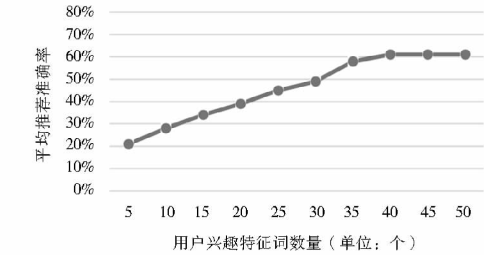


图2　P＝5时平均推荐准确率变化趋势

当Q＝40时，令P依次取5、10、15、20、25、30、35、40、45、50得到如图3所示的8名用户的平均推荐准确率。随着P值的不断增大，平均推荐准确率逐渐提高，当P为30、40、50时平均推荐准确率基本不再发生变化，因此取P＝30，此时改进模型已达到最优的推荐准确率。

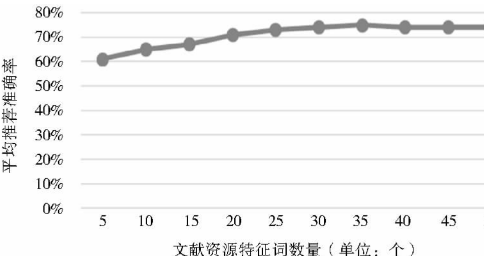


图3　Q＝40时平均推荐准确率变化趋势

取Q＝40,P＝30，观察改进后模型随着推荐文献数量从30至70每次增加5的不同情况下的查准率P、召回率R、F1值及运行时间的变化情况。在相同的实验环境与数据条件下，对改进前的基于向量空间模型的推荐进行实验，利用TF-IDF文本表示方法对文本数据进行量化，将文本数据以权值向量的形式表示出来，即将用户兴趣与文献资源分别表示成空间向量模型的形式，通过计算向量的相似度为目标用户推荐文献。此外，本文还利用相同的数据对基于LDA主题模型的推荐进行实验，利用LDA主题模型对实验数据进行训练，得到每篇文献的主题分布概率，抽取文献在主题上的概率分布作为特征向量，然后计算文献间的相似度，以此为用户进行推荐。

图4给出了改进模型、向量空间模型及LDA主题模型推荐的查准率P的变化情况。由图4可以看出，

随着推荐文献数量的增加,3 种推荐模型的查准率 P 都在下降,当推荐文献数量高于 45 时,查准率并无明显差异。在实际应用中最为重要的是推荐的排名靠前的文献数量,改进模型在推荐文献数量小于 45 时查准率有明显优势,所以本研究所提出的改进模型与实验中的向量空间模型与 LDA 主题模型相比具有更高的应用价值。

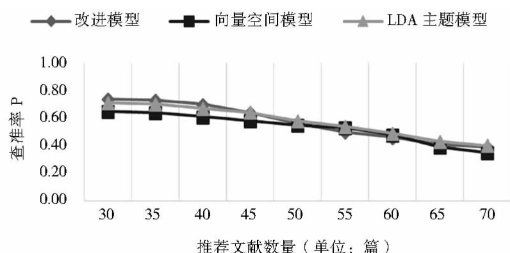


图 4　3 种推荐模型的查准率 P 随推荐文献数量增加的变化情况

图 5 给出了 3 种推荐模型随着推荐文献数量从 30 增加到 70 每次增加 5 的不同情况下召回率 R 的变化情况。从图 5 中可以看出,改进模型在推荐文献数量小于 45 时召回率 R 优于其他两种推荐模型。

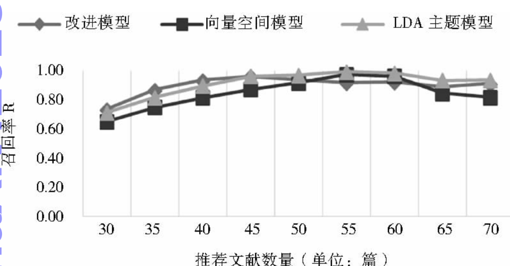


图 5　3 种推荐模型的召回率 R 随推荐文献数量增加的变化情况

图 6 给出了 3 种推荐模型随着推荐文献数量从 30 增加到 70 每次增加 5 的不同情况下的 F1 值的变化情况,同样由图 6 可以看出当推荐文献数量小于 45 时,改进模型明显优于其他两种模型。

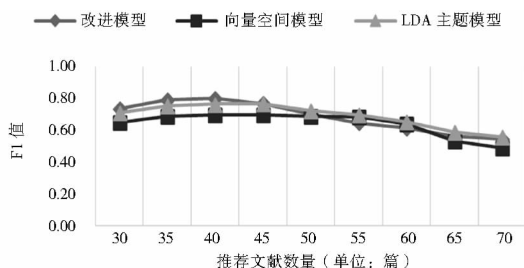


图 6　3 种推荐模型的 F1 值随推荐文献数量增加的变化情况

图 7 给出了 3 种推荐模型随着推荐文献数量从 30 增加到 70 每次增加 5 的不同情况下的运行时间对比图,改进模型的平均运行时间为 6.19 秒,向量空间模型的平均运行时间为 11.45 秒,LDA 主题模型的平均运行时间为 7.12 秒。相比较于向量空间模型的推荐,改进模型的的运行效率提升了 45.93%,且优于 LDA 主题模型的 7.12 秒。

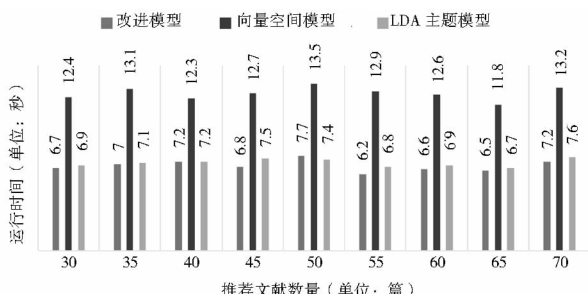


图 7　3 种推荐模型运行时间对比

实验结果表明,当 Q 为 40,P 为 30 时,改进模型的推荐准确率达到最优。改进模型在推荐文献数量小于 45 时,其查准率、召回率与 F1 值明显优于其他两种推荐模型,而用户更看重的是排名靠前的文献,因此,改进模型在现实中具有更高的应用价值。另外在模型运行时间复杂度方面,相比较于改进前的向量空间模型,改进模型的运行效率提升了 45.93%,且优于 LDA 主题模型。

5　结语

针对基于向量空间模型的推荐中在表示文本时维度过大的问题,本研究提出利用余弦值 r 与匹配度值 Sim 相结合的方法对原有模型进行改进,结合已有的 TF-IDF、word2vec 等技术构建了基于内容过滤的个性化推荐模型。该模型通过筛选出权重较大的特征词及特征词向量来计算 r 与 Sim,通过对改进模型、向量空间模型及 LDA 主题模型进行的实验结果表明,改进模型的推荐准确率较好运行效率高,在现实中具有更高的应用价值。未来可将改进模型应用到更大的数据集上,提高模型的稳定性和运行效率,本文的研究成果为高校图书馆个性化推荐服务提供了新的思路。

参考文献:

[1] 何胜,熊太纯,柳益君,等. 基于 Spark 的高校图书馆文献推荐方案及实证研究[J]. 图书情报工作,2017,61(23):129 – 137.

[2] 刘旭晖. 融合主题多样性与影响力的科技文献推荐算法研究[J]. 情报理论与实践,2017,40(12):134 – 138.

[3] 刘伟,刘柏嵩,王洋洋. 海量学术资源个性化推荐综述[J]. 计算机工程与应用,2018,54(3):30 – 39.

[4] 王超,吕俊生. 国内外学术信息推荐方法研究进展[J]. 情报杂

志,2013, 32(9):142－147.

[5] 阮光册,夏磊. 推荐系统的发展与公共图书馆个性化信息服务探讨[J]. 图书馆,2016(2):94－99.

[6] BEEL J, GIPP B, LANGER S, et al. Research-paper recommender systems: a literature survey[J]. International journal on digital libraries, 2015, 17(4):1－34.

[7] PHILIP S, MUSA E P. A paper recommender system based on the past ratings of a user[J]. International journal of advanced computer technology (IJACT), 2014,3(6):41－46.

[8] 黄震华, 张佳雯, 张波,等. 语义推荐算法研究综述[J]. 电子学报, 2016, 44(9):2262－2275.

[9] 徐勇,司凤山,吴延辉, 等. 基于概念泛化的科技文献推荐算法[J]. 图书情报工作,2012,56(21):101－108.

[10] 吕学强,王腾,李雪伟, 等. 基于内容和兴趣漂移模型的电影推荐算法研究[J]. 计算机应用研究,2018,35(3):717－720,802.

[11] 安悦,李兵,杨瑞泰,等. 基于内容的热门微话题个性化推荐研究[J]. 情报杂志,2014,33(2):155－160.

[12] 丁德红,方逵,王娟, 等. 基于内容过滤推荐的农业信息推荐模型研究[J]. 湖南农业大学学报(自然科学版),2013,39(6):683－687,562.

[13] 雷凯,刘树波,李丹, 等. 实时路况制约下基于内容的兴趣点推荐[J]. 计算机工程,2017,43(10):147－152.

[14] LIU L, LECUE F, MEHANDJIEV N. Semantic content-based recommendation of software services using context[J]. ACM transactions on the web (TWEB), 2013, 7(3): 1－20.

[15] DELDJOO Y, ELAHI M, CREMONESI P, et al. Content-based video recommendation system based on stylistic visual features[J]. Journal on data semantics, 2016, 5(2): 99－113.

[16] LEE S, CHOI K, SUH Y. A personalized trustworthy seller recommendation in an open market[J]. Expert systems with applications, 2013, 40(4): 1352－1357.

[17] LIU N H. Comparison of content-based music recommendation using different distance estimation methods[J]. Applied intelligence, 2013, 38(2): 160－174.

[18] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009, 29(s1):167－170.

[19] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer science, 2013:arXiv:1301.3781.

[20] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in neural information processing systems. Nevada:Curran Associates Inc. 2013: 3111－3119.

[21] LAI S, LIU K, HE S, et al. How to generate a good word embedding[J]. IEEE intelligent systems, 2016, 31(6):5－14.

[22] 王飞,谭新. 一种基于 Word2Vec 的训练效果优化策略研究[J]. 计算机应用与软件,2018,35(1):97－102,174.

作者贡献说明:

耿立校:确定研究命题及研究框架,撰写论文及定稿;
晋高杰:撰写论文、模型构建、实例分析及论文修改;
李亚函:论文的逻辑思路、语言组织及撰写指导;
孙卫忠:实证分析部分用户数据的收集;
马士豪:实证分析部分数据的预处理。

Research on Personalized Recommendation of University Library Literature Resources Based on Improved Content-based Filtering Algorithm

Geng Lixiao[1] Jin Gaojie[1] Li Yahan[1] Sun Weizhong[1,2] Ma Shihao[1]
[1] School of Economics and Management, Hebei University of Technology, Tianjin 300401
[2] Hebei University of Technology Library, Tianjin 300401

Abstract: [**Purpose/significance**] In content-based filtering recommendation, the problem of dimensionality disaster is easily caused when the vector space model (VSM) is used to represent text. This paper proposes a method that combines the cosine value r and the matching value Sim to improve the original model. [**Method/process**] based on literature resources and user interests the word vectors of feature words with large weight were selected, and then the cosine value r is calculated by the formula, and the matching value Sim is further calculated based on the corresponding feature words weights as the basis for recommending literature to the target user. And it uses the data from the Hebei University of Technology Library to conduct experiments on the improved model, vector space model and LDA topic model, and finally uses the evaluation index of precision rate, recall rate, F1 and running time to analysis the experimental results of the three models. [**Result/conclusion**] The experimental results show that the improved model presented in this paper has higher application value and operation efficiency compared with the vector space model and LDA topic model.

Keywords: content-based recommendation matching value Sim recommendation model empirical analysis